Resolução de ambiguidades

Eric Laporte

Capítulo 2

Resolução de ambiguidades

Éric Laporte

Resolução de ambiguidades[1]

Éric Laporte

A resolução de ambiguidades, intrínsecas a qualquer língua natural, é uma operação básica cujos resultados são úteis para quase todos os processamentos de textos escritos, inclusive para alguns dos processamentos mais simples. Contudo, não é uma tarefa fácil. Examinaremos, utilizando exemplos precisos, quais os problemas que surgem na elaboração de sistemas de resolução de ambiguidades lexicais. Para avaliarmos o potencial dos métodos, teremos em conta as melhorias decorrentes da resolução de ambiguidades em relação aos principais processamentos de texto escrito.

1. Etiquetagem lexical de textos

O processamento de textos escritos não pode ser feito sem informações linguísticas relativas às palavras. Para dispor de tais informações rápida e convenientemente, os programas informáticos costumam associá-las às próprias palavras dos textos sob a forma de etiquetas lexicais. A etiqueta lexical de uma palavra reúne, portanto, todas as informações disponíveis sobre ela e úteis ao processamento, desde a própria forma que consta no texto até aos dados gramaticais, morfológicos, sintácticos ou semânticos, dependendo da natureza do processamento. Uma etapa primordial consiste em segmentar o texto, identificar as unidades mínimas e representá-las por etiquetas. Essa etapa é designada como análise lexical ou etiquetagem lexical.

Os meios técnicos para associar as informações lexicais às palavras podem ser classificados em dois tipos, conforme as informações provêm de um dicionário electrónico ou são decorrentes de informações presentes no texto.

A etiquetagem lexical por dicionário electrónico é simples: o programa procura as palavras do texto num dicionário que associa etiquetas a todas as palavras da língua. Este método foi largamente posto à prova nos anos 90 e dá os resultados mais fiáveis, desde que o conteúdo do dicionário esteja conforme à realidade da língua e suficientemente perto da exaustividade. Nas línguas flexivas, como a maioria das línguas europeias, trata-se de dicionários de palavras flexionadas, nos quais o número de entradas é maior do que num dicionário convencional, onde os verbos constam só no infinitivo (Ranchhod, «O uso de dicionários e de autómatos finitos na representação lexical das línguas naturais» neste volume). As línguas altamente flexivas, como por exemplo o polaco, têm vários milhões de palavras flexionadas. Mesmo assim, existem dicionários razoavelmente próximos da exaustividade, que podem ser comprimidos em ficheiros com cerca de 1 Mb, e permitem a etiquetagem lexical de milhares de palavras por segundo. O sistema INTEX oferece ferramentas eficazes de compressão de dicionários e de etiquetagem de textos (Silberztein, 1994). Neste capítulo, utilizaremos as convenções habituais do INTEX para a formalização das etiquetas lexicais: assim, <*activo*,A:fp> representa o adjectivo *activo* no feminino plural, isto é, *activas*.

Os métodos de etiquetagem lexical sem dicionário foram postos em prática por numerosos sistemas nos anos 80 e 90. Apoiam-se em informações presentes no texto, como o sufixo das palavras e o seu contexto. Por exemplo, muitas palavras terminadas em

[1] Este trabalho foi parcialmente financiado pela União europeia no quadro do projecto Copernicus 621 "GRAMLEX".

-ivas são adjectivos no feminino plural. Esta regra atribui correctamente a etiqueta <A:fp> à palavra *activas* na frase:

As empresas mais activas no mercado cresceram

Muitas palavras precedidas por *alguns dos* são substantivos no masculino plural. Esta regra, quando aplicada à frase:

Alguns dos filmes premiados só serão exibidos no outono

associa correctamente a *filmes* a etiqueta <N:mp>. Regras semelhantes são elaboradas automaticamente por treino, baseado na contagem de frequências em amostras de textos etiquetados ou não. São conhecidos vários tipos de treino. O próprio princípio é uma aproximação, pois palavras com a mesma parte final podem ter propriedades inteiramente diferentes, por exemplo o substantivo *cartazes* e o verbo *trazes*. Contudo, é a unica solução para etiquetar palavras desconhecidas.

2. Ambiguidades lexicais

As dificuldades residuais da etiquetagem lexical são devidas à omnipresença de ambiguidades lexicais em todas as línguas. Há ambiguidade lexical entre dois elementos linguísticos distintos quando se escrevem exactamente da mesma forma. Algumas ambiguidades lexicais resultam de imperfeições dos sistemas ortográficos das línguas. Por exemplo, nas duas frases:

É proibido colher flores
Use a sua colher

as duas formas *colher*, um verbo e um substantivo que se pronunciam de modo diferente, são idênticas pela simples razão de a ortografia portuguesa não distinguir sistematicamente as duas vogais tónicas *e*, o que constitui uma dificuldade para os programas de transcrição fonética de textos escritos (Viana, «Síntese de fala» neste volume). Este tipo de ambiguidade lexical é mais ou menos abundante, e depende das imperfeições dos sistemas ortográficos; teoricamente, não existiria se a ortografia de uma dada língua fosse suficientemente informativa.

Todavia, as ambiguidades lexicais abrangem uma realidade bem mais extensa do que algumas curiosidades e coincidências. Examinemos vários exemplos de palavras ambíguas. Nas seguintes frases, as duas ocorrências de *forma* correspondem, respectivamente, a um verbo e a um nome sem que haja entre elas qualquer diferença fonética:

Um especialista não se forma facilmente
O contrato foi redigido de forma a satisfazer ambas as partes

Embora um leitor humano não detecte a menor ambiguidade, a necessidade de associar duas etiquetas diferentes a um verbo e a um nome, como etapa prévia ao reconhecimento da estrutura sintáctica das frases, por exemplo, é evidente. Dois termos de uma ambiguidade lexical podem ter uma relação etimológica, como *forma*, ou não, como *decorar*:

(1) *Os alunos decoraram o soneto*
(2) *As vendedoras decoraram a vitrina*

Podem até ser dois elementos de um mesmo paradigma:

O gato foge
Por favor, foge agora

Neste caso, as etiquetas respectivas diferem por traços flexionais: o presente do indicativo ou o imperativo, nestes exemplos. Muitas ambiguidades lexicais estabelecem-se entre palavras que têm uma relação etimológica estreita e pertencem à mesma categoria gramatical, mas possuem acepções distintas:

(3) *Ela marcou o livro com um pedacinho de papel*
(4) *Este livro marcou a minha adolescência*

Em geral, as propriedades sintácticas das várias acepções são diferentes: no caso dos verbos, o número de complementos essenciais, as preposições associadas, as tranformações sintácticas aplicáveis, a distribuição dos nomes que podem ser colocados na frase na posição de sujeito ou de complemento podem diferir. Ora, tais informações sintácticas, necessárias para o reconhecimento da estrutura da frase, devem ser incluídas nas etiquetas lexicais.

Outro tipo frequente de ambiguidades lexicais envolve palavras compostas:

(5) *A mesa redonda sobre a política de saúde foi cancelada*
(6) *A mesa redonda é pequena, usemos a outra*

Na frase (5), *mesa redonda* é um nome composto, e, como tal, uma unidade mínima de processamento. A etiqueta deverá, pois, conter informações sintácticas relacionadas, por exemplo, com a forma e o tipo de complementos desta unidade. Em (6), *mesa* e *redonda* são unidades distintas que constituem um grupo nominal livre. A tradição subestima a importância das palavras compostas nos textos, mas estudos recentes (Senellart, 1999) avaliam em mais de metade a percentagem de texto constituída por palavras compostas. Além disso, o conteúdo técnico de um texto, inclusive a maioria dos termos técnicos, encontra-se mais nessa parte do que na outra. Em comparação com a tecnologia padrão geralmente aceite pela comunidade de linguística computacional, a operação de etiquetagem lexical deve aplicar-se a unidades muito mais complexas do que as palavras simples. Etiquetar só as palavras simples seria limitar-se a uma parte superficial da língua.

A existência de ambiguidades lexicais tem consequências técnicas. O papel de um dicionário electrónico é o de garantir que as etiquetas de todas as palavras estão disponíveis; portanto, em caso de acepções múltiplas, a mera consulta do dicionário fornece as informações lexicais relativas a todas as acepções, e não apenas as informações requeridas. Nessa altura, as ambiguidades lexicais estão explicitamente representadas por listas de etiquetas; é preciso resolvê-las para terminar a etiquetagem lexical e chegar à situação na qual cada palavra está representada só pela etiqueta, ou pelas etiquetas, correctas. Aliás, uma ambiguidade lexical pode perfeitamente tornar ambígua a frase onde se encontre; é o caso de (1), em que os alunos podem ter decorado a página do soneto com desenhos nas margens. Trata-se de um exemplo de frase cuja ambiguidade é representável ao nível da etiquetagem lexical, mas que não pode ser resolvida.

O caso da etiquetagem lexical sem dicionário é diferente: é possível gerar várias etiquetas representativas de várias análises, mas é impossível garantir que a etiqueta correcta de cada palavra se encontre entre as etiquetas produzidas. Além disso, o reconhecimento das palavras compostas sem dicionário é um problema quase insolúvel, pois a maioria delas obedecem a todas as regras habituais de concordância morfológica e de sintaxe superficial, como *mesa redonda*, *levar em conta*, *a respeito de*, sendo, portanto, impossíveis de detectar sem dispor de listas completas. De agora em diante, limitar-nos-emos a ter em conta a etiquetagem por dicionário.

Os dicionários electrónicos podem estar mais ou menos longe da exaustividade, por descreverem um número diferente de palavras, ou de acepções de palavras. Quando o número de acepções descritas cresce, as ambiguidades lexicais também crescem, o que pode parecer um obstáculo ao processamento no caso de acepções raras, como *galo* (gaulês). Para determinadas aplicações, este argumento é uma razão para limitar a extensão da descrição lexical, mas o trabalho fundamental de descrição dos dados linguísticos não se pode limitar a um tipo de aplicações: a descrição de palavras raras e de acepções raras é útil para outras aplicações. Um raciocínio segundo o qual a solução para evitar o aparecimento de ambiguidades lexicais consistiria em restringir a descrição lexical, seria o mesmo que impedir o desenvolvimento do transporte aéreo a fim de evitar a construção de aeroportos.

3. Granularidade

As informações lexicais contidas nos dicionários, e depois associadas às palavras dos textos que vão ser automaticamente processados, podem ser mais ou menos pormenorizadas, dependendo da natureza e da quantidade da informação tida em conta aquando da descrição lexical. Por exemplo, etiquetas lexicais reduzidas à indicação da categoria gramatical: verbo, substantivo, adjectivo... são pouco informativas. A seguinte frase é um exemplo desta etiquetagem minimalista, com abreviações clássicas:

(7) *Esperava<V> a<Det> sua<Det> chegada<N> à tarde<Adv>*

A riqueza das informações lexicais está ligada à importância numérica do conjunto de etiquetas lexicais: as categorias gramaticais formam um conjunto de cerca de 15 elementos, mas se se tomarem em conta mais informações, este número só pode crescer. Designaremos este parâmetro qualitativo como a granularidade da descrição, uma vez que, à medida que esta cresce, cada etiqueta descreve menos palavras, mas fá-lo de maneira mais informativa.

A granularidade da descrição aumenta com a inclusão de novas informações. Os conjuntos de etiquetas mais usados para o inglês compreendem essencialmente, além da categoria gramatical, traços flexionais: tempo verbal, número, pessoa... O conjunto do Penn Treebank totaliza 36 etiquetas (Marcus *et al.*, 1993), o do British National Corpus 58 (Leech *et al.*, 1994), e o de Greene e Rubin 86 (1971). Nas línguas neolatinas, que são mais flexionadas do que o inglês, os atributos equivalentes perfazem de 80 a 120 etiquetas. Acrescentemos os traços flexionais à frase (7), considerando apenas uma das possíveis interpretações:

(8) *Esperava<V:I3s> a<Det:fs> sua<Det:fs> chegada<N:fs> à tarde<Adv>*

Podemos ainda incluir nas etiquetas a forma canónica de cada palavra flexionada, por exemplo o infinitivo no caso dos verbos:

(9) *Esperava<esperar,V:I3s> a<o,Det:fs> sua<seu,Det:fs> chegada<chegada,N:fs> à tarde<à tarde,Adv>*

Para limitar o número de etiquetas sem perder informação, as formas canónicas podem ser abreviadas, aproveitando a redundância entre a forma flexionada e a forma canónica: com essa convenção, *esperava<esperar*,V:I3s> é substituído por *esperava<2r*,V:I3s>. O número de etiquetas distintas atinge uma ordem de grandeza de 1000 no caso do francês. Os dicionários morfológicos da rede de laboratórios RELEX, que estão integrados no sistema INTEX, usam conjuntos de etiquetas dest Ainda e tipo.

Por agora, só incluímos nas etiquetas lexicais informações gramaticais, morfológicas e flexionais. Ainda assim, a granularidade da descrição lexical variou bastante. A representação das ambiguidades lexicais depende do sistema de etiquetas lexicais: a palavra *esperava* não era representada como ambígua em (7):

esperava<V>

A mesma palavra torna-se ambígua logo que tomamos em conta os traços flexionais, pois pode tratar-se da primeira, da segunda (*você*) ou da terceira pessoa:

esperava<V:I1s>
esperava<V:I4s>
esperava<V:I3s>

Assim, a ambiguidade lexical, representada por meio de um conjunto de etiquetas lexicais, cresce automaticamente com a granularidade. Contudo, as estatísticas sobre textos indicam que esse aumento é moderado enquanto as informações se limitarem ao nível gramatical, morfológico e flexional. Calculámos a ambiguidade lexical média de uma amostragem de textos do francês em 1,63 etiquetas por palavra com um conjunto de etiquetas do tipo de (7), e em 1,99 com um conjunto do tipo de (9). O aumento das ambiguidades lexicais é limitado (22%) em comparação com o enriquecimento da informação proporcionado pela substituição de conjuntos de etiquetas.

Do ponto de vista das aplicações, é preciso avaliar a adequação dos conjuntos de etiquetas aos requisitos das aplicações. O processamento mais simples de textos escritos é a detecção de erros lexicais, isto é, de palavras que não fazem parte do vocabulário, como *quilometro*: esta operação não necessita de etiquetas. Contudo, certas aplicações, já mais ambiciosas, requerem uma etiquetagem lexical do texto, mas podem ser consideradas aceitáveis apesar de não darem resultados exaustivos. Por exemplo, os utilizadores de sistemas de detecção de erros não lexicais estão geralmente cientes da dificuldade da tarefa, vêem esses sistemas como uma ajuda à releitura dos textos e não confiam neles para apontar exaustivamente os erros. O mesmo se passa em relação à busca automática de informações em textos e à indexação de documentos. Os textos seleccionados, no primeiro caso, e as entradas de índice escolhidas, no outro, podem conter elementos não desejados e elementos desejados podem faltar, sem pôr em perigo a utilização dos sistemas, pois os desejos do utilizador são definidos de modo aproximativo e os resultados estão destinados a serem tratados manualmente. Para esta categoria de aplicações, as informações contidas em etiquetas do tipo de (9) permitem certamente a obtenção de resultados interessantes, em comparação com o padrão de resultados actual.

Outras aplicações, ainda mais exigentes, requerem muito mais informações lexicais: a geração de fala a partir de texto escrito livre, a tradução, e todas as outras aplicações que envolvem uma análise sintáctica aprofundada, isto é, um reconhecimento da estrutura das frases e dos constituintes das frases: orações, predicados, complementos. Esse reconhecimento não pode ser automatizado sem informações lexicais específicas.

No caso de verbos e de outras formas predicativas, como *marcar, dar um pulo, estar de acordo, ser devedor, dar de caras*, o número de complementos essenciais e as preposições associadas são indispensáveis para identificar o elemento ou os elementos predicativos da frase e os seus eventuais argumentos: sujeito e complementos. Além disso, a frase pode ser o resultado da aplicação de uma ou várias operações sintácticas: passiva, inversões, reduções, omissões, pronominalizações... Ora, os especialistas em sintaxe sabem que nem todas as transformações sintácticas são aplicáveis a todas as formas predicativas, e essa informação é essencial para o reconhecimento da estrutura sintáctica. Também, a distribuição dos nomes que podem ser colocados na frase na posição de sujeito ou de

complemento depende de cada forma predicativa: o conhecimento, mesmo incompleto, dessa distribuição é fundamental para discriminar hipóteses durante a análise sintáctica (por exemplo, entre (3) e (4)).

Uma vez que se trata de propriedades lexicais, a sua formalização tem de ser feita em extensão. O léxico-gramática (M. Gross, 1994) é o quadro formal e metodológico natural para a realização de tal estudo. Examinemos um exemplo simples de informação linguística indispensável à análise sintáctica: a posição pré-nominal ou pós-nominal dos adjectivos. Essa informação é lexical, no sentido em que depende de cada adjectivo e não segue regras gerais (Carvalho, em preparação). O adjectivo *simpático* pode ser pré- ou pós-nominal, o adjectivo *político* é pós-nominal:

uma simpática oferta
uma oferta simpática
**uma política questão*
uma questão política

Cada uma das palavras *adulto* e *analfabeto* é ambígua entre adjectivo e nome:

Propomos um curso para adultos analfabetos

Nesta frase, só a combinação das ambiguidades das categorias gramaticais gera quatro análises para *adultos analfabetos*; a informação de que o adjectivo *adulto* é exclusivamente pós-nominal elimina correctamente a análise *adulto*<A:mp> *analfabeto*<N:mp>. O impacto quantitativo deste tipo de regra em relação à resolução de ambiguidades lexicais em português e em francês (Garrigues, 1997) é importante.

Com a inclusão de informações exploráveis na análise sintáctica, a riqueza informativa e a granularidade das etiquetas crescem consideravelmente, uma vez que a descrição formal dessas propriedades sintácticas envolve a separação das acepções dos verbos, como em (3) e (4), dos adjectivos e de outros elementos predicativos. Nesta altura, a granularidade do sistema de descrição tem muito pouco que ver com os sistemas de etiquetagem sem dicionário que constituem o padrão actual. As propriedades sintácticas que evocámos são difíceis de manusear por sistemas cujo funcionamento esteja baseado em estatísticas de frequências. Por exemplo, a posição pré-nominal ou pós-nominal dos adjectivos franceses não está correlacionada com marcas ortográficas, como a presença de sufixos, pode depender das acepções de uma mesma forma, variando muitas vezes livremente para uma determinada acepção, o que dificulta a tarefa de obter essa informação por generalização automática a partir de exemplos tirados de uma amostragem de textos.

Uma consequência do aumento da granularidade da descrição é o aumento das ambiguidades lexicais: cada palavra com várias acepções associadas a propriedades sintácticas diferentes pode *a priori* estar representada por etiquetas correspondentes, combinando-se essas ambiguidades com as ambiguidades flexionais que já ilustrámos. Ainda não há estudos sobre o número médio de etiquetas por palavra, mas uma ordem de grandeza de 10 é plausível. Porém, o aumento das ambiguidades lexicais não advém de um defeito do sistema de formalização, é antes um reflexo da complexidade do problema. Um conjunto de etiquetas com uma alta granularidade é uma ferramenta pesada do ponto de vista técnico, mas pelas mesmas razões que fazem com que uma escavadora tenha de ser pesada: um terreno não se pode aplanar com uma colher de chá.

No que respeita à quantidade de ambiguidades lexicais em geral, é importante notar que ela só pode ser medida, tecnicamente, através de uma contagem do número de etiquetas lexicais por palavra, número que depende, entre outros factores, da granularidade do conjunto de etiquetas. Assim, a comparação entre sistemas de taxas de erro, de taxas de redução, isto é, a proporção de ambiguidades resolvidas pelos respectivos processamentos,

e de outros resultados numéricos de sistemas baseados em conjuntos de etiquetas diferentes, não é de modo algum significativa.

4. Delimitação do problema e dos objectivos

Examinámos vários exemplos de ambiguidades lexicais e vimos que o fenómeno afecta qualquer texto, praticamente qualquer frase, até as mais simples. Além disso, dependendo da granularidade do sistema de descrição, a quantidade de ambiguidades e o número de etiquetas podem crescer notavelmente, o que é desejável que aconteça devido à complexidade das aplicações mais interessantes que motivam o estudo do problema.

Neste momento, não podemos deixar de questionar se é possível resolver todas as ambiguidades lexicais, e a que custo. A resposta é clara. Para determinadas frases, a resolução de todas as ambiguidades lexicais obriga ao reconhecimento completo da estrutura sintáctica. Nas seguintes frases, *parecer* é, respectivamente, verbo e nome:

(10) *Tinha a necessidade de parecer favorável à ideia*
(11) *Lembrava-se de um exemplo de parecer favorável ao réu*

O contexto imediato é idêntico nas duas frases, tanto à esquerda: <Det:s> <N:s> <*de*,Prep>, como à direita: <*favorável*,A:s> <*a*,Prep> <*o*,Det:s> <N:s>. É preciso conhecer as propriedades dos nomes *necessidade* e *exemplo* para resolver a ambiguidade. O complemento nominal de *necessidade* é uma completiva infinitiva:

(12) *Ele tem necessidade de decidir*

A forma nominal do complemento não aceita o determinante vazio no singular:

Ele tem necessidade de uma decisão rápida
(13) **Ele tem necessidade de decisão rápida*
Ele tem necessidade de decisões rápidas

O nome *exemplo* não selecciona uma completiva infinitiva:

(14) **Este texto é um exemplo de descrevermos os nossos resultados*

O complemento nominal de *exemplo* obedece a restrições diferentes sobre a determinação:

Este texto é um exemplo de um relatório favorável
(15) *Este texto é um exemplo de relatório favorável*
Estes textos são exemplos de relatórios favoráveis

As restrições sintácticas (13) e (14), se estiverem formalizadas e disponíveis no dicionário, excluem respectivamente <*parecer*,N:ms> de (10) e <*parecer*,V:W> de (11); no entanto, as construções (12) e (15) justificam respectivamente a escolha de <*parecer*, V:W> em (10) e de <*parecer*, N:ms> em (11).

Existem exemplos de frases ambíguas em relação às quais a resolução total das ambiguidades lexicais necessita da análise completa das estruturas sintácticas envolvidas. A ambiguidade da frase:

Encheu a colher de chá

provém da ambiguidade lexical do nome *colher de chá*. Numa das interpretações, o nome composto *colher de chá* é o complemento directo do verbo *encher*, e o segundo complemento essencial do verbo está ausente. Nas outras interpretações, o nome *colher* é o complemento directo e *de chá* é um segundo complemento. Como esse complemento é facultativo, as duas intepretações são possíveis, mas é preciso fazer uma análise completa

da frase, a partir das propriedades sintácticas do verbo, para chegar a essa conclusão. Aliás, uma análise análoga da seguinte frase leva à resolução da mesma ambiguidade:

Encheu a colher de chá de sal

A posição pré-nominal ou pós-nominal dos adjectivos fornece outros exemplos em que a resolução total das ambiguidades lexicais requer uma análise sintáctica mais aprofundada do que a que foi antes referida.

Assim, a determinação correcta das etiquetas lexicais associáveis às palavras pode depender do reconhecimento da estrutura sintáctica global da frase. Trata-se de uma dependência circular, já que o reconhecimento das estruturas sintácticas se baseia nas informações lexicais incluídas nas etiquetas. Uma das dificuldades inerentes à análise sintáctica é precisamente essa dependência circular.

A observação de que a resolução exaustiva das ambiguidades lexicais depende, em geral, da análise sintáctica global, afecta radicalmente a natureza do problema. A etiquetagem lexical correcta é um subproduto da análise sintáctica. Deste modo, a resolução de ambiguidades fica sem objecto e sem solução, e teoricamente desaparece enquanto problema distinto.

Contudo, um objectivo de resolução *parcial*, ou redução, das ambiguidades lexicais, que não necessita de uma análise sintáctica completa, é menos ambicioso e mais realista. Tal objectivo refere-se a um processamento distinto, concebido para eliminar análises inválidas antes da análise sintáctica, e é por vezes designado como filtragem dos resultados de uma etiquetagem lexical por dicionário electrónico. Os resultados da etiquetagem lexical, antes dessa filtragem, constituem um conjunto de análises do texto ou da frase, e cada uma dessas análises está representada como uma sequência de etiquetas. Acredita-se que tal processamento de filtragem pode ajudar a análise sintáctica do texto, ao limitar o número de análises concorrentes de uma frase e a complexidade das informações transmitidas ao analisador sintáctico, fazendo assim com que chegue ao mesmo resultado em menos tempo (Milne, 1986). Os sistemas de etiquetagem, ou de análise sintáctica, que funcionam por eliminação de análises têm sido referidos com a designação de reducionistas (Voutilainen e Tapanainen, 1993).

O interesse de uma tal filtragem depende, porém, do conteúdo algorítmico da componente de análise sintáctica e, nomeadamente, da relação entre as suas entradas e a sua velocidade.

Se o tempo gasto pela análise sintáctica for proporcional ao número de análises das entradas, uma filtragem rápida das entradas tem grande probabilidade de aumentar a velocidade global da operação, pois o número de análises de uma frase cresce exponencialmente com o número médio de etiquetas por palavra, e, portanto, aumenta muito rapidamente com as ambiguidades lexicais.

O tempo gasto pelos algoritmos modernos de análise sintáctica não depende, no entanto, do número de análises das entradas, mas sim da complexidade da estrutura que representa essas análises. Essa estrutura pode ser um autómato finito. Ora, quando um autómato finito representa um conjunto de sequências, a complexidade do autómato não depende directamente do número de sequências. Quando só há uma sequência, o autómato é sempre pequeno; mas, quando o processamento de filtragem elimina algumas das sequências, a complexidade do autómato pode aumentar ou diminuir. Por isso, a complexidade do autómato não pode ser usada como uma ferramenta quantitativa para medir a eficácia da filtragem. Treoricamente, a redução das ambiguidades lexicais pode aumentar ou diminuir a velocidade da análise sintáctica.

De acordo com as nossas experiências em relação ao francês (Laporte, Monceaux, 1999), a redução das ambiguidades lexicais é rápida, proporcionando em geral uma forte

diminuição da complexidade do autómato que representa as análises concorrentes da frase. É plausível, pois, que a aplicação de bons sistemas de resolução de ambiguidades lexicais aumente a velocidade da análise sintáctica. Todavia, esta hipótese deverá ser verificada empiricamente quando se dispuser de analisadores sintácticos satisfatórios, capazes de explorar o conteúdo de etiquetas lexicais razoavelmente informativas. É teoricamente possível que a filtragem parcial das ambiguidades lexicais torne, de facto, o processamento globalmente mais lento em vez de mais rápido, e que a utilidade de tal componente desapareça completamente a longo prazo.

Mas não é ainda o caso: pelo contrário, tudo indica que a existência de bons sistemas de redução das ambiguidades lexicais facilitará o desenvolvimento de analisadores sintácticos.

Definiremos, assim, a resolução de ambiguidades lexicais como um processamento aplicado às análises resultantes da etiquetagem lexical, cujo objectivo é eliminar o maior número possível de análises inválidas, utilizando os meios mais simples e mais rápidos possíveis.

Esta definição tem várias consequências importantes.

Em primeiro lugar, o carácter aproximativo, estatístico, deste objectivo caracteriza-o como um problema de aplicação, por oposição a problemas mais fundamentais como a descrição lexical e a análise sintáctica, que têm objectivos sistemáticos. Assim, a avaliação de um sistema de resolução de ambiguidades lexicais não é de natureza teórica, é antes uma questão puramente empírica. O aspecto pouco teórico do problema pode diminuir a motivação para o estudar, mas o seu enorme interesse aplicativo é uma compensação. Além do mais, o carácter aproximativo do objectivo não implica de modo algum que a tarefa seja fácil ou que a elaboração da solução possa ser o resultado de um trabalho aproximativo: veremos mais tarde a razão deste aparente paradoxo.

Em segundo lugar, a redução das ambiguidades lexicais só tem sentido em combinação com outro processamento, como a análise sintáctica, e, neste caso, observa-se um estreito acoplamento entre a resolução das ambiguidades lexicais e a análise sintáctica, no sentido de não ser possível automatizar a primeira operação sem ter em conta o modo de automatização da segunda, e que dois sistemas informáticos que efectuem as respectivas operações de modo compatível não podem ser modificados de forma independente. Por exemplo, os dois sistemas devem usar o mesmo conjunto de etiquetas lexicais e basear-se no mesmo tipo de análises. Ao ser o resultado da redução das ambiguidades lexicais transmitido ao analisador sintáctico, a fiabilidade deste depende da fiabilidade daquele. Retomaremos este assunto adiante.

Em terceiro lugar, a divisão das operações entre uma etapa de resolução de ambiguidades lexicais e uma etapa de análise sintáctica faz com que o que não for feito numa tenha que ser feito na outra. Por exemplo, cabe ao analisador sintáctico resolver as ambiguidades que restem no fim da primeira etapa. O sentido desta divisão reside no facto de que a primeira etapa se limita a meios simples e rápidos. A organização do sistema global, porém, precisa de definir esses limite de modo mais preciso. De acordo com a prática geral, pode dizer-se que a resolução das ambiguidades lexicais:

- não reconhece sistematicamente os constituintes,
- não insere símbolos de fronteira (com excepção do de frase),
- não gera novas análises além das decorrentes directamente da etiquetagem lexical, e
- não representa explicitamente nos resultados as transformações sintácticas aplicadas,

A análise sintáctica, no entanto, não pode deixar de recorrer, de um modo ou de outro, a cada um destes meios técnicos. Em consequência destas limitações impostas à redução de ambiguidades lexicais para a manter simples e rápida, a análise do contexto só pode ser

local, o que não significa que se limite apenas a uma palavra de cada lado. Consideremos, por exemplo, a ambiguidade de *capitais*:

(16) *Há dois tipos de capitais europeias, segundo a minha experiência: as medievais e as modernas.*

A acepção financeira corresponde ao masculino: <*capital*,N:mp>, e a acepção geográfica ao feminino: <*capital*,N:fp>. Uma regra de concordância local entre nome e adjectivo adjacentes pode aproveitar o facto de *europeias* estar no feminino para escolher correctamente a etiqueta <*capital*,N:fp>. Se o adjectivo *europeias* não existisse na frase:

Há dois tipos de capitais, segundo a minha experiência: as medievais e as modernas.

as únicas informações susceptíveis de resolver a ambiguidade de *capitais* encontrar-se-iam nos grupos nominais incompletos à direita dos dois pontos. O reconhecimento da relação entre *capitais* e esses grupos nominais necessitaria, entre outras coisas, do reconhecimento do complemento *segundo a minha experiência*, um objectivo que sai do âmbito da resolução de ambiguidades lexicais e pertence à análise lexical. Pode, assim, considerar-se que o contexto necessário à resolução da ambiguidade de *capitais* não é, neste caso, local.

Em quarto lugar, é preciso distinguir as noções de silêncio (em que uma análise válida é eliminada) e de ruído (em que uma análise inválida é conservada). Silêncio e ruído são duas formas de desvio do objectivo da redução de ambiguidades lexicais. Este objectivo, tal como foi acima definido, visa diminuir o ruído sempre que possível; sabemos que isso nem sempre é completamente conseguido, pelo que a resolução das ambiguidades residuais cabe normalmente à análise sintáctica. Consideremos agora as consequências da eliminação de uma análise válida. O resultado da etapa de filtragem é processado por outra componente, que pode ser um analisador sintáctico. Já verificámos que a fiabilidade desta componente depende da fiabilidade da filtragem de ambiguidades: a consequência inevitável da eliminação de uma análise válida é o fracasso da análise sintáctica de uma frase inteira. Ora, um analisador sintáctico, por sua vez, é um programa destinado a servir de componente essencial a sistemas de tradução, de geração de fala a partir de textos escritos, ou de outras aplicações nas quais a fiabilidade dos resultados é um parâmetro importante. A geração de resultados fiáveis num tempo aceitável é o principal objectivo de um analisador sintáctico, e a etapa de filtragem é introduzida só para acelerar o processamento. Manter o silêncio a zero durante a etapa de filtragem das ambiguidades lexicais é, portanto, um objectivo incondicional e prioritário. O objectivo de manter o ruído ao nível mais baixo possível é aproximativo e secundário. Devido a esta ordem de prioridades, é aconselhável muita prudência no uso de aproximações aquando da redução de ambiguidades lexicais.

5. Dados requeridos pela resolução de ambiguidades

A resolução automática de ambiguidades lexicais envolve análise e reconhecimento do contexto gramatical, a fim de verificar restrições locais. Por exemplo, o estudo da frase (16) ilustra um caso de aproveitamento de restrições que se manifestam na concordância entre substantivo e adjectivo. Trata-se, em geral, de restrições distribucionais, gramaticais e combinatórias sobre as sequências de palavras ou de etiquetas lexicais. Estas restrições, devidamente formalizadas, constituem os dados linguísticos do sistema, e podem ser designadas por gramáticas de resolução de ambiguidades.

As questões relativas à proveniência e aos métodos de elaboração desses dados são muito debatidas. Eles podem ser elaborados quer por linguistas quer por aprendizagem automática.

A primeira solução é a mais antiga (Harris, 1962; Klein, Simmons, 1963; Greene, Rubin, 1971). O processo de descrição e de formalização de restrições gramaticais pode ser artesanal ou mais ou menos industrializado, mas requer uma reflexão analítica não trivial. Consideremos o exemplo de *parecer*, ambíguo entre nome e verbo:

(17) *Calou-se para não parecer indiscreta*

Os indícios de que, nesta frase, *parecer* é verbo são muito locais: a presença de *não* e do adjectivo no feminino *indiscreta*. Esta observação pode ser formalizada como:

(18) Quando está imediatamente precedido por *não*, *parecer* é um verbo.

Esta restrição gramatical elimina correctamente a etiqueta nominal de *parecer* em (17). A dificuldade reside na determinação do nível de generalidade adequado. Uma restrição pouco geral, como (18), aplica-se raramente e resolve poucas ambiguidades. Podemos generalizá-la, uma operação também natural e intuitiva:

(19) Qualquer palavra imediatamente precedida por *não* é um verbo.

Esta nova versão ainda se aplica correctamente a (17), mas elimina incorrectamente a etiqueta <*aumento*,N:ms> de *aumento* em:

O não aumento dos impostos foi uma medida eleitoralista

(cf. Ranchhod, 1989). Uma restrição demasiado geral aplica-se a casos inadequados e pode eliminar análises válidas. Isso põe em perigo a fiabilidade do sistema de filtragem de ambiguidades lexicais e do eventual analisador sintáctico, afastando o sistema do seu objectivo prioritário de manter o silêncio em zero. Para determinar um nível de generalidade aceitável, os responsáveis pela descrição e formalização das restrições dispõem de dois métodos:

- a pesquisa de exemplos e de contra-exemplos em textos[2], e
- a construção directa de exemplos e sobretudo de contra-exemplos.

Estes dois métodos são complementares. O primeiro, embora seja parcialmente automatizável, não se pode substituir ao segundo. Consideremos um exemplo do francês, retirado de Senellart (1999):

(20) Numa sequência com a forma (*a, as, avions*) <ADV> <V:K>, a primeira palavra é uma forma do verbo *avoir*,

em que <V:K> representa os particípios. Esta restrição está concebida para eliminar a ambiguidade de *a*, *as* e *avions*, três formas do verbo *avoir* homógrafas de substantivos. Elimina correctamente a etiqueta de nome em:

Nous avions également popularisé une technologie

e nenhum erro de aplicação foi detectado num ano do jornal *Le Monde*. Mesmo assim, existem contra-exemplos óbvios e naturais:

Nous ne pilotons que des avions complètement révisés

em que (20) elimina incorrectamente a etiqueta de nome.

O exemplo ilustrado por (18)-(19) é particularmente simples, uma vez que recorre a um contexto muito local. Na prática, é frequentemente preciso, como em (20), ter em conta um contexto ligeiramente mais alargado, à direita, à esquerda ou dos dois lados. O autor das descrições e da formalização das restrições deve pensar em todos os contextos possíveis de uma dada palavra. Repetimos que, por definição, está privado das ferramentas

[2] O testar o sistema em textos é uma forma de automatizar a pesquisa.

de análise sintáctica, que seriam o reconhecimento sistemático dos grupos nominais e de outros constituintes.

A segunda solução para a construção dos dados linguísticos de sistemas de redução de ambiguidades baseia-se numa generalização (ou indução) automática a partir de textos etiquetados ou não (Marshall, 1983; Benello *et al.*, 1989; Merialdo, 1994), e, por vezes, a partir de outros dados como esquemas de regras pré-definidos (Brill, 1992). Trata-se de uma solução inerentemente aproximada, orientada pelo tratamento dos casos que ocorrem com muita frequência nos textos.

O resultado da generalização automática pode apresentar-se sob a forma de regras legíveis, ou de dados numéricos ilegíveis. No último caso, o comportamento do sistema de resolução de ambiguidades não está definido, isto é, o resultado da aplicação do sistema a um determinado texto pode ser conhecido, testando o sistema, mas não pode ser predito. Por outras palavras, não há garantias quanto aos resultados, e nomeadamente o objectivo prioritário de conservar todas as análises válidas está fora de alcance.

Além disso, etiquetas lexicais muito informativas e contextos constituídos por mais do que uma palavra são difíceis de explorar pelos métodos baseados em estatísticas de frequência. De facto, a generalização automática pode ser vista como a exploração de um espaço abstracto cujo volume depende, entre outras coisas, do número de etiquetas lexicais, da extensão dos contextos, e da dimensão da amostra de textos que serve de base à exploração. Aliás, o aumento do primeiro destes três parâmetros pode motivar um aumento dos outros dois. Até agora, a possibilidade de testar o método com etiquetas lexicais de alta granularidade está limitada pela quantidade de cálculo requerido pela exploração deste espaço. Já vimos, por exemplo, que a separação das acepções dos adjectivos é difícil de tratar dentro de um tal quadro.

De agora em diante, consideraremos apenas os problemas relativos à filtragem de ambiguidades lexicais, feita a partir de dados obtidos através de descrições e formalizações directamente realizadas por linguistas.

6. Exemplos de restrições gramaticais

As restrições distribucionais ou gramaticais formalizadas são a matéria principal de um sistema de filtragem parcial de ambiguidades lexicais. É importante avaliar as dificuldades que surgem na elaboração desses dados, e, por isso, ilustrá-las-emos com abundantes exemplos. Cada restrição gramatical descrita separadamente é por vezes designada por regra, que evoca em geral a presença de uma rede de regras e excepções. Sempre que o usarmos, não pressuporemos, porém, a existência de um sistema de relações regra/excepção. Examinaremos mais tarde os problemas de manutenção, característicos deste tipo de sistemas.

A definição dos nossos objectivos, tendo em conta o carácter primordial da fiabilidade de um sistema de filtragem, cujos resultados são explorados por outros sistemas, aponta como prioritário o objectivo de garantir que o processo conserve sempre as análises válidas. É um objectivo muito difícil de alcançar. Consideremos as ambiguidades das palavras: *completa*, verbo ou adjectivo; e *programa* e *visita*, verbos ou nomes:

(21) *Ele completa o programa com uma visita à universidade*

As palavras gramaticais *ele* e *uma* são, respectivamente, indícios de que *completa* é um verbo, e *visita* um nome. As restrições combinatórias sobre o emprego dessas palavras podem ser formalizadas assim:

(22) *ele, ela, eles, elas* não podem preceder imediatamente um adjectivo.

(23) *um, uma* não podem preceder imediatamente um verbo.

Estas regras eliminam correctamente as etiquetas[3] <*completo*.A:fs> e <*visitar*.V:P3s> em (21), mas possuem contra-exemplos:

Achei a ideia desenvolvida por ele completa e inovadora
Aquele que quase comprou uma passa por aqui todos os dias

Nestas frases, (22) e (23) eliminam incorrectamente <*completo*.A:fs> e <*passar*.V:P3s>. Um contra-exemplo é suficiente para desqualificar uma restrição que elimina análises válidas. A dificuldade em evitar este tipo de erro é inerente ao problema. Por um lado, o responsável da descrição e da formalização das restrições gramaticais deve conhecer todos os contextos gramaticais observáveis das formas processadas, e tomá-los em conta, o que não acontece em (22) e (23). Pelo outro, deve reconhecer localmente um contexto suficiente nas frases processadas, sem reconhecimento prévio dos limites de sintagmas, embora esse contexto possa ser ambíguo. Por exemplo, a restrição (22) pode ser melhorada tomando em conta um contexto mais vasto à esquerda, o que diminui a sua generalidade. Se um limite de frase explícito ou uma conjunção subordinativa ocorrerem imediatamente à esquerda de <*ele*,Pro>, (22) aplica-se melhor: o facto sintáctico subjacente é a presença de um limite de oração, mas a formalização das restrições gramaticais não se pode referir aos limites de orações, pois a filtragem das ambiguidades lexicais é feita antes da análise sintáctica que trata, entre outras coisas, do reconhecimento desses limites.

A formalização de restrições gramaticais para redução das ambiguidades lexicais apresenta outro aspecto desagradável: a impossibilidade de completar a descrição das restrições. Não é de surpreender, já que o objectivo de uma descrição sintáctica completa sai do quadro definido. Contudo, dois casos, ainda que sintacticamente muito parecidos, podem não ser processáveis pela mesma restrição. Voltemos, por exemplo, à regra (22) que se aplica correctamente imediatamente depois de um limite de frase explícito, como em (21). Sintacticamente, a adjacência entre o pronome e o verbo é um pormenor contingente da estrutura, pois entre *ele* e *completa* pode ser inserido um constituiente, sem que isso altere a frase de base:

Ele, de repente, acostumado a tomar decisões no último momento, completa o programa com uma visita à universidade

Todavia, a regra (22) já não se aplica agora, e não pode ser adaptada de modo a que isso aconteça, uma vez que tal significaria o reconhecimento completo do complemento acrescentado.

Estas dificuldades inerentes tornam a formalização das restrições gramaticais uma tarefa difícil, por vezes, frustrante. Até se pode considerar que uma tal tarefa é irrealista, pelo facto de que qualquer gramática de resolução de ambiguidades lexicais tem certamente contra-exemplos. Também se pode pensar que os (poucos) sistemas existentes já tentaram resolver essas dificuldades de todos os meios ao dispor e alcançaram os melhores resultados possíveis. Contudo, essas opiniões não se apoiam em factos verificáveis. Pelo contrário, pensamos que as dificuldades inerentes ao problema são bem conhecidas, mas que as soluções não foram ainda investigadas sistematicamente. Tal investigação pode orientar-se em duas direcções complementares: por um lado, as análises linguísticas subjacentes ao processamento total; por outro lado, o formalismo de descrição das restrições gramaticais. Examinaremos estes dois assuntos sucessivamente.

7. As análises linguísticas subjacentes

[3] Conforme as notações habituais do INTEX, as formas incluídas nas etiquetas dentro dos símbolos < > são formas canónicas.

A redução de ambiguidades lexicais insere-se sempre num processamento global, cuja finalidade é a atribuição de uma análise linguística a frases de textos escritos, ou de várias análises, no caso de frases ambíguas. Este objectivo, por sua vez, é um pré-requisito para certos tipos de processamento de texto. As análises linguísticas são descrições formais que vão desde a representação dos elementos mínimos do texto até à formalização da estrutura sintáctica das frases. A questão das análises a serem atribuídas às frases é, obviamente, um assunto fundamental, embora seja pouco debatido na literatura da especialidade. De facto, as várias componentes informáticas que concorrem para o processamento devem referir-se à mesma análise linguística subjacente, o que cria uma interdependência entre essas componentes.

A descrição formal da estrutura sintáctica de uma frase passa pelas fases seguintes:

- as unidades elementares são identificadas no dicionário electrónico e as etiquetas lexicais correspondentes associadas às palavras, inclusive às palavras compostas; nesta altura, o que chamamos de análise da frase corresponde a uma sequência de etiquetas lexicais;
- esta análise atravessa o filtro de resolução parcial de ambiguidades;
- durante a análise sintáctica, são reconhecidos os constituintes e as transformações sintácticas, das quais a frase é um exemplo.

As três etapas do processo são a consulta do dicionário, a redução das ambiguidades lexicais e a análise sintáctica. As três componentes informáticas que as realizam apoiam-se num conjunto de dados linguísticos, respectivamente no dicionário electrónico, nas gramáticas de resolução de ambiguidades lexicais, e na descrição sintáctica da língua. A sua interrelação requer que estes conjuntos de dados estejam baseados nas mesmas análises. Acontece, por vezes, que podem ser concebidas várias análises de uma dada frase, e que podem surgir problemas quando são escolhidas análises diferentes para a elaboração dos dados linguísticos, ou quando a análise escolhida é formalizada de formas diferentes.

Consideremos a frase seguinte:

(24) *Aquele livro é bom, mas esse é pouco interessante*

O sujeito de *é pouco interessante* pode ser analisado de duas formas. A primeira solução é considerar que *esse* é, nesta frase, um pronome; o dicionário electrónico deve, portanto, descrevê-lo como ambíguo entre determinante, <*esse*.DET:ms>, e pronome, <*esse*.PRO:ms>, para além de um ou vários substantivos <*esse*.N:ms>. A etiqueta de determinante é concebida para situações do tipo:

(25) *Esse livro é bom*

e a de pronome para:

(26) *Esse é bom*

A outra solução consiste em representar os dois casos com a mesma etiqueta, e formalizar a relação sintáctica entre (25) e (26) como uma transformação de apagamento do substantivo. Note-se, aliás, que, qualquer que seja a etiquetagem lexical da palavra, a formalização dessa relação para uma análise sintáctica aprofundada é indispensável por outras razões, nomeadamente a reconstituição completa do sujeito de (26). Na segunda solução, o dicionário representa *esse* como ambíguo entre um determinante <*esse*.DET:ms> e um ou vários substantivos <*esse*.N:ms>. As duas soluções são aplicáveis a outros determinantes: *aquele, este, algum, seu, dois, três...*, e são equivalentes: a diferença entre elas não aparece como uma verdadeira divergência de análise linguística, antes como um pormenor de formalização.

Contudo, o manuseamento de conceitos linguísticos num sistema informático necessita de formalização, e uma vez formalizadas as análises linguísticas subjacentes, um mero detalhe de formalização basta para as tornar distintas. Uma vez que a representação das palavras em discussão é diferente no dicionário, as gramáticas de redução de ambiguidades lexicais e o analisador sintáctico devem ser coerentes com a solução escolhida. Na primeira solução, *esse* está representado como mais ambíguo do que na segunda. Consideremos, então, a seguinte restrição gramatical:

(27) Um verbo não pode ocorrer imediatamente à direita de um determinante.

A restrição (27) pode ser correcta se *esse* for analisado como pronome em (26), e, de facto, é coerente com a primeira solução, mas é incorrecta se *esse* for analisado como determinante. Se for adoptada a segunda solução, a ambiguidade entre <*esse*.DET:ms> e <*esse*.PRO:ms> desaparece, bem como a necessidade de (27); portanto, a segunda solução simplifica simultaneamente o dicionário e a gramática de redução de ambiguidades. É interessante observar os sintomas que podem aparecer em caso de discrepância entre o dicionário e a gramática de redução das ambiguidades.

Se o dicionário for coerente com a primeira solução e a gramática com a segunda, aparece uma ambiguidade lexical entre <DET> e <PRO>, que não é eliminada pela gramática; pode considerar-se que se trata de uma ambiguidade artificial, pois é um artefacto resultante da incoerência entre duas componentes dos dados linguísticos.

Se, pelo contrário, o dicionário for coerente com a segunda solução e a gramática com a primeira, a regra (27) aplica-se e elimina a análise em que *esse* é determinante, conservando só a análise ou as análises incorrectas em que *esse* é substantivo. Portanto, a análise correcta é eliminada pela filtragem.

A origem do problema reside na presença de duas representações formais da mesma construção gramatical. Outro exemplo prototípico da mesma situação é a possibilidade de representar a sintaxe de uma construção, quer como livre, quer como fixa. A sintaxe de uma construção é dita livre se for considerada como uma combinação de palavras simples, como *vinho bom*, e fixa se constituir uma palavra composta, como *vinho branco*. Uma forma pode ser ambígua entre uma construção livre e uma palavra composta, como *mesa redonda*. Trata-se de uma ambiguidade de natureza lexical, pois o acesso às propriedades lexicais da palavra composta pressupõe o seu reconhecimento ao nível da análise lexical. Por vezes, as ambiguidades lexicais podem ser resolvidas localmente, como é provavelmente o caso de *a fim* em:

Mudou de roupa a fim de se sentir mais à vontade,

em que é resolvida a favor da palavra composta. O problema que nos interessa aqui aparece quando se pode hesitar entre duas representações, uma livre e uma fixa, para a mesma construção. A sintaxe da acepção tipográfica de *maiúsculo* (Senellart, 1999) é um exemplo dessa situação. Todas as ocorrências da acepção tipográfica de *maiúsculo* correspondem às seguintes formas:

- combinações do tipo *N maiúsculo*, em que *N* apenas pode ser preenchido ou pelos substantivos *letra, abecedário, alfabeto, caracter, inicial* ou pelos nomes das letras do alfabeto, *a, b*, etc.:

**No texto aparece um nome de pessoa maiúsculo*

Estas formas não são associáveis a frases do tipo *N é maiúsculo*:

**O caracter inicial de um nome próprio é maiúsculo*

O substantivo *letra* pode sempre aparecer numa paráfrase dessas formas: *letra inicial maiúscula, alfabeto de letras maiúsculas*, etc.

- o substantivo *maiúscula* é sempre equivalente a *letra maiúscula*;

- a forma *N estar em maiúscula(s)* é parafraseável por *N estar escrito em letra(s) maiúscula(s)*.

A sintaxe destas expressões é tão restrita que pode ser representada por uma gramática local (M. Gross, *Construção de gramáticas locais e autómatos finitos*, neste volume). A distribuição dos nomes nas formas *N maiúsculo* é tão reduzida e específica que se pode considerar que todas essas formas constituem uma pequena família de nomes compostos, e o substantivo *maiúscula* como uma redução do nome composto *letra maiúscula*. Sendo estes compostos representados por etiquetas lexicais, não é preciso recorrer a uma etiqueta adicional de adjectivo para a acepção tipográfica de *maiúsculo* (no português do Brasil, há, pelo menos, outra acepção adjectival diferente desta).

Uma segunda solução, mais conforme à tradição gramatical e lexicográfica, seria a de considerar o substantivo *maiúscula* como um substantivo simples, e as formas *N maiúsculo* como combinações livres de nomes e de um adjectivo <*maiúsculo*, A> independente. A sintaxe restrita da combinação deveria, portanto, ser representada nos dados do analisador sintáctico, nomeadamente a distribuição do *N*, a impossibilidade de empregar o «adjectivo» de forma predicativa ou atributiva, e a redução por apagamento do substantivo *letra*.

A diferença entre as duas soluções reside no tratamento, nos dicionários ou nas gramáticas, dos dados sobre a sintaxe semi-fixa de uma família de expressões. De qualquer modo, esses dados devem ser formalizados, mas podem aparecer quer no dicionário (as gramáticas locais podem ser consideradas como parte do dicionário), quer na gramática da sintaxe da língua. Na primeira solução, a forma *maiúsculo* não é representada como ambígua, pois a etiqueta lexical depende do substantivo associado à esquerda. Na segunda solução, a forma é sistematicamente representada como ambígua (nome e adjectivo), devendo essa ambiguidade ser resolvida, o que complica o processamento sem qualquer vantagem visível.

As consequências de uma discrepância entre os modelos seguidos para o dicionário e para a gramática de filtragem das ambiguidades lexicais dependem do tipo de incoerência de que se trate. Se o dicionário representar as formas como fixas, e a gramática de resolução de ambiguidades lexicais as considerar como livres, a resolução da ambiguidade entre nome e adjectivo a favor do nome não tem efeitos nocivos, pois a etiqueta de adjectivo a eliminar nem consta no resultado da análise lexical. Se, pelo contrário, o dicionário representar as formas como livres e a gramática de resolução de ambiguidades lexicais as considerar como fixas, a ambiguidade entre nome e adjectivo introduzida pela análise lexical não é resolvida: a discordância entre os dois conjuntos de dados linguísticos provoca, portanto, uma ambiguidade artificial.

Apresentaremos um terceiro exemplo, o do género dos adjectivos nas línguas neolatinas. Alguns adjectivos variam em género, como *fixo/fixa*, outros não, como *fiel*. A distinção do masculino e do feminino aparece como uma informação gramatical relevante no caso de *fixo*, mas como um factor de ambiguidade lexical artificial no caso de *fiel*, pois leva ao manuseamento de duas etiquetas <*fiel*.A:ms> e <*fiel*.A:fs>. O mesmo pode ser dito de certos determinantes, pronomes e nomes.

Deste ponto de vista, o sistema flexional das línguas neolatinas pode ser formalizado de dois modos. A primeira solução, tradicional, considera o género como uma categoria

relevante para todos os adjectivos, e representa *fiel* como flexionalmente[4] ambíguo. Na segunda solução, o género é relevante para alguns adjectivos, e irrelevante para os outros. A escolha entre as duas soluções tem que ver com os três conjuntos de dados em discussão: o dicionário, que representa *fiel* com uma etiqueta ou duas; a gramática de resolução de ambiguidades lexicais, que ambiciona resolver a ambiguidade flexional de *fiel* ou não; e a gramática que formaliza a sintaxe da língua. Se *fiel* for representado como ambíguo, a representação formal de todos os adjectivos é mais uniforme, o que simplifica a formalização das concordâncias entre nome e adjectivo. A escolha de uma solução envolve a avaliação global da situação. Nem será preciso dizer que a coexistência dos dois modelos no mesmo processamento não pode deixar de provocar um verdadeiro desastre, embora se trate de um detalhe de formalização.

Os três exemplos: a relação entre determinantes e pronomes, a sintaxe fixa de *maiúsculo*, o género dos adjectivos, evidenciam a interdependência entre os elementos da descrição linguística. A construção do dicionário electrónico, das gramáticas de resolução de ambiguidades lexicais, e até a construção do conjunto de etiquetas lexicais, devem ter em conta a mesma análise linguística subjacente ao processamento total, incluindo a análise sintáctica. Os problemas devidos a eventuais incoerências ou incompatibilidades entre elementos de descrição linguística não invalidam o próprio esquema geral que adoptámos, nomeadamente a repartição dos dados e do processamento em componentes distintas e compatíveis. Essas dificuldades práticas são bem conhecidos dos investigadores que têm uma actividade descritiva efectiva no domínio, e só salientam a complexidade da descrição linguística em grande escala. Os sintomas visíveis que resultam de uma discrepância entre elementos de descrição linguística formal podem ser o aparecimento de silêncio (a eliminação de análises que se desajavam conservar), de ambiguidades artificiais ou ainda o aumento do ruído.

O papel das análises linguísticas subjacentes à descrição formal é obviamente fundamental. Referiremos duas consequências que dependem da escolha dessas análises:

- a simplicidade do processamento global; escolher soluções descritivas globalmente simples é um objectivo importante, devido à complexidade inerente das línguas quando se tem o léxico em conta;

- a quantidade de ambiguidades lexicais representadas, medidas pelo número de etiquetas lexicais por palavra, número que depende das etiquetas associáveis às palavras no quadro das análises linguísticas escolhidas, inclusive do número de etiquetas das palavras gramaticais do dicionário; portanto, resultados numéricos relativos a sistemas baseados em análises linguísticas diferentes, ou formalizadas diferentemente, não são comparáveis.

Os dois aspectos: a simplicidade da descrição formal e a quantidade de ambiguidades lexicais das palavras gramaticais estão, aliás, interligados. No exemplo da relação entre determinante e pronome por apagamento do substantivo (25)-(26), a solução mais simples é caracterizada por um número de etiquetas menor, portanto por menor ambiguidade, porque o problema é considerado sintáctico. Esse exemplo é prototípico de uma situação frequente com palavras gramaticais. Quando a coordenação entre os vários conjuntos de dados linguísticos é máxima, a quantidade de ambiguidades lexicais diminui dado que os elementos linguísticos são apenas representados uma vez.

Contudo, esse efeito não chega para eliminar completamente as ambiguidades lexicais de todas as palavras gramaticais, nem para tornar inútil a atribuição de etiquetas lexicais. Certas palavras gramaticais apresentam ambiguidades sem nenhuma ligação

[4] A combinação de duas etiquetas sob a forma de <*fiel*.A:ms:fs> ou de <*fiel*.A:mfs>, por exemplo, é uma variante notacional desta primeira solução.

sincrónica em relações sintácticas do tipo de (25)-(26). Por exemplo, *se* pode ser conjunção ou pronome:

(28) *Gostaria desse se pudéssemos conservá-lo*
Nem se consegue ler

A atribuição de pelo menos duas etiquetas lexicais distintas <*se*.CONJ> e <*se*.PRO:3> à forma *se* tem uma utilidade prática: estabelecer uma comunicação entre várias restrições gramaticais. Por exemplo, em (28), uma restrição gramatical pode eliminar a etiqueta <*se*.PRO:3> dado que o verbo *pudéssemos* está na primeira pessoa do plural, e outra restrição pode utilizar a outra etiqueta, <*se*.CONJ>, como uma marca de limite de oração. É conveniente formalizar as duas restrições gramaticais separadamente, pois nada têm a ver entre si, e só a etiquetagem lexical de *se* possibilita o uso do resultado da primeira pela segunda. As etiquetas lexicais de palavras gramaticais ainda são, pois, um instrumento conveniente para a redução eficaz das ambiguidades lexicais. Além disso, várias palavras gramaticais são homógrafas de palavras lexicais: o pronome e determinante *este* é homógrafo de uma variante do nome *leste*, a forma *são* tanto pode ser verbo como adjectivo...

A construção de gramáticas de resolução de ambiguidades lexicais é uma actividade importante na medida em que contribui para um trabalho persistente de elaboração progressiva da gramática das línguas, de formalização e aplicação à análise automática dos textos. A longo prazo, esse poderá ser mesmo o seu aspecto mais interessante.

8. **Necessidade de formalismos para a resolução de ambiguidades lexicais**

A resolução de ambiguidades lexicais pressupõe:

- a descrição de restrições gramaticais durante a construção do sistema, e
- a filtragem automática das análises conformes às restrições, durante a aplicação do sistema ao processamento de textos.

A descrição e a automatização da filtragem não podem ser levadas à prática sem um formalismo de descrição que defina as convenções de formalização das restrições gramaticais e a interpretação da descrição.

A descrição linguística, em geral, pode ser mais ou menos formal. A falta de formalização pode tornar a descrição aproximativa, incompleta, incoerente ou inaplicável. Um alto nível de formalização não constitui, contudo, por si só uma garantia sobre a adequação do conteúdo; pelo contrário, se significar uma multiplicidade de detalhes de formalização, constitui um obstáculo à descrição. A questão do formalismo de descrição condiciona, pois, o sucesso do empreendimento.

A automatização da filtragem, embora seja uma questão mais informática, não é independente da interpretação da descrição formal, pois a filtragem consiste em eliminar as análises que não satisfazem as restrições gramaticais. Uma função importante do formalismo de descrição é a de garantir a possibilidade de construir uma filtragem automática conforme às restrições descritas.

Vamos discutir e ilustrar quais as razões que impõem o uso de um formalismo, ou de formalismos, para reduzir as ambiguidades lexicais.

Na ausência de um formalismo definido, é impossível definir uma noção de restrição gramatical correcta, pois a interpretação das restrições será sempre aproximativa. Consideremos a seguinte restrição gramatical:

(29) Imediatamente depois de um <N:ms>, um <A> singular está no masculino.

Tendo em conta que as línguas naturais são muito ambíguas, a restrição (29) pode ser correcta ou incorrecta. Se a interpretação for no sentido de que ela elimina as análises nas

quais uma etiqueta lexical <N:ms> é imediatamente seguida por uma etiqueta lexical <A:fs>, pode ter uma certa relevância; resolve correctamente, por exemplo, a ambiguidade de género do adjectivo na seguinte frase:

Reconheceu o caminho habitual

Se a interpretação for a de que (29) se aplica depois de qualquer palavra que possa ser um <N:ms>, revela-se incorrecta, pois elimina, incorrectamente, a análise de *pertinente* na seguinte frase:

Sempre se tem achado pertinente a inclusão desse tipo de documentos

devido à ambiguidade de *achado* entre particípio e substantivo. O problema aqui advém do facto de a restrição gramatical não estar expressa de forma precisa. Não sendo clara a natureza exacta da restrição gramatical, não é possível determinar se ela é correcta ou incorrecta.

A função principal do formalismo de descrição é, então, muito simples: garantir que a interpretação das restrições descritas seja clara e precisa, de tal forma que se possa definir uma noção de restrição correcta (conforme às análises linguísticas subjacentes), e que a filtragem das restrições correctas seja automatizável.

É lógico e legítimo tentar imaginar o formalismo mais simples possível. Por exemplo, um formalismo mínimo, quase uma ausência de formalismo, consistiria em formalizar restrições gramaticais sob a forma de meros conjuntos de sequências gramaticais, exactamente como se fossem gramáticas locais descrevendo famílias de formas. A concordância entre substantivo e adjectivo poderia ser:

(30) <N:ms> <A:ms> + <N:mp> <A:mp> + <N:fs> <A:fs> + <N:fp> <A:fp>

Os tempos compostos dos verbos com o auxiliar *ter* seriam representados como:

<*ter*.V> <V:Kms>

Infelizmente a ideia é demasiado simples. Descrições de formas não definem explicitamente restrições gramaticais; por outras palavras, não eliminam análises de forma explícita. Deste ponto de vista, (30) é até menos explícita do que (29). Para interpretar uma descrição de formas do tipo de (30) como uma restrição gramatical, sem ambiguidade, seria preciso definir convenções de intepretação. Várias convenções seriam possíveis; corresponderiam a interpretações distintas, isto é, a restrições gramaticais distintas. Convenções de interpretação diferentes teriam vantagens e inconvenientes para o autor da descrição; para serem usadas, deveriam ser realizadas informaticamente por programas diferentes. Tratar-se-ia, portanto, de formalismos distintos, inevitavelmente mais complexos do que a ideia inicial.

O exemplo precedente salienta que a especificação explícita dos efeitos da aplicação de uma gramática a um texto faz parte de um formalismo de resolução de ambiguidades lexicais. De facto, um formalismo não pode ser usado satisfatoriamente sem tal especificação, que prediz que tipo de análises são eliminadas pela aplicação de gramáticas expressas pelo formalismo. A especificação serve de referência comum aos autores das gramáticas, que descrevem restrições, e aos autores do programa informático, que realiza a filtragem. A utilidade da especificação torna-se evidente no caso de ocorrência de um erro numa dessas duas partes do sistema, quer nos dados linguísticos, quer no programa. Quando um erro é descoberto, a especificação é usada para determinar a qual das partes diz respeito: por exemplo, a filtragem realizada pelo programa não está conforme à especificação decidida. Enquanto o erro é corrigido, a especificação mantém-se como referência comum, não sendo a outra parte perturbada. Na ausência de especificação, as

duas partes do sistema são perturbadas por falta de referência: o comportamento que servia de referência anteriormente está errado, e o novo comportamento não é conhecido enquanto o erro estiver a ser corrigido. O recurso a especificações é sistemático na construção de sistemas informáticos complexos: é uma forma de organizar a comunicação entre partes complexas de um sistema. Evidentemente, a noção de especificação não resolve todos os problemas: as especificações podem mudar, o que perturba o sistema na sua totalidade, mas isso acontece mais raramente do que os erros de programação ou de descrição linguística.

Assim, a finalidade de um formalismo de resolução de ambiguidades lexicais é a de garantir que as restrições gramaticais nele expressas sejam definidas de forma precisa e sem ambiguidade. A primeira qualidade de um formalismo é provavelmente a sua simplicidade, mas, como vimos, o problema em discussão não é tão simples que permita reduzir o formalismo a zero.

Tratando-se de formalismos descritivos, a comodidade que traz aos autores de descrições é a principal orientação para definir critérios de qualidade:

- a simplicidade e a legibilidade,
- a predictibilidade dos efeitos da aplicação das gramáticas a um texto,
- a possibilidade de organizar a descrição em componentes pequenas e relativamente independentes entre si,
- a possibilidade de realizar um programa informático eficaz que leve à prática a filtragem especificada.

Vejamos três outras orientações que, de acordo com a nossa experiência de construção e de utilização de gramáticas de resolução de ambiguidades, dizem respeito à qualidade do formalismo de descrição.

8.1. *Etiquetas de pontuação e etiquetas-variáveis*

As etiquetas lexicais completas, que descrevem palavras determinadas, como <*esse*.DET:ms>, são úteis à resolução de ambiguidades, porque podem ser usadas na formalização de restrições distribucionais ou gramaticais específicas dessas palavras. Contudo, essas restrições dependem geralmente do contexto gramatical, e é frequente ser preciso tomar em conta ou reconhecer um contexto local para definir as suas condições de aplicação. O reconhecimento do contexto pode envolver o uso de etiquetas que descrevem sinais de pontuação, ou que representam categorias de palavras, como categorias gramaticais, por exemplo <N>, que representa qualquer substantivo. Todos os exemplos de restrições acima compreendem, por exemplo, etiquetas deste último tipo, excepção feita à restrição (18). Tais etiquetas podem ser designadas como etiquetas-variáveis, por representarem um conjunto de etiquetas lexicais clássicas. As convenções de formalização das etiquetas, incluindo as etiquetas-variáveis, são um elemento importante do formalismo de descrição das restrições gramaticais. As convenções habituais possibilitam a combinação de categorias gramaticais:

- com um ou vários traços flexionais: <N:m>, que representa qualquer substantivo masculino, <V:1s>, qualquer verbo na primeira pessoa do singular,
- ou com uma forma canónica: <*sadio*.A>, <*sadio*.A:f>, etc.

Outros tipos de etiquetas-variáveis podem ser úteis, como, por exemplo, as etiquetas negativas: <!*ser*!*estar*!*ter*.V> para representar qualquer verbo, com exclusão de *ser, estar* e *ter*.

8.2. *Delimitação do poder expressivo do formalismo*

Um formalismo pode ter maior ou menor poder expressivo, o que quer dizer que oferece mais ou menos meios técnicos para exprimir restrições gramaticais de tal ou tal tipo. Geralmente, os autores de descrições consideram qualquer limitação do poder expressivo do formalismo utilizado como um obstáculo ao seu trabalho de elaboração e de generalização a partir das suas intuições.

Contudo, paradoxalmente, uma limitação do poder expressivo do formalismo pode facilitar esse trabalho de formalização. Já usámos, por exemplo, uma convenção de formalização de etiquetas que proíbe o uso de um tipo de etiquetas-variáveis inteiramente natural: as etiquetas reduzidas a uma palavra, como *são*, que representam qualquer etiqueta lexical associável à forma citada. No exemplo, há pelo menos duas: <*ser*.V:P3p> e <*são*,A:ms>. Segundo esta convenção, todas as etiquetas lexicais incluem pelo menos a categoria gramatical, como <*são*,A:ms>, <A:ms>, <*são*,A>, excepto a etiqueta universal que representa qualquer palavra. Uma gramática escrita para resolver uma ambiguidade lexical relacionada com a forma *são*, não se podendo referir a esta forma por intermédio da etiqueta-variável *são*, recorrerá necessariamente a etiquetas lexicais que incluem a categoria gramatical, como <*ser*.V:P3p> e <*são*,A:ms>. Esta convenção apresenta vantagens por facilitar o trabalho de descrição formal, embora constitua uma limitação ao poder expressivo do formalismo.

- Em caso de dúvida a respeito das etiquetas lexicais associadas a uma forma na análise linguística subjacente ao sistema, a convenção incita o descritor a referir-se ao dicionário electrónico e a ter em conta o seu conteúdo.

- Os contextos relevantes às duas etiquetas (verbo e adjectivo, no exemplo de *são*) não têm qualquer razão para serem parecidos. Quando o autor da regra descreve os contextos relevantes em relação a uma delas, não precisa de ter em conta a outra, nem mesmo a ambiguidade da forma. Em geral, a descrição das restrições gramaticais torna-se mais cómoda se se puderem ignorar completamente as eventuais ambiguidades dos elementos descritos. O paradoxo é só aparente, dado que a existência de ambiguidades é a razão pela qual se empreende a construção das gramáticas, mas são os contextos distribucionais e sintácticos específicos dos elementos linguísticos o que importa descrever e não as ambiguidades. Aliás, durante a filtragem, as análises são processadas como se fossem totalmente distintas.

- Dado que as restrições gramaticais são expressas a partir de etiquetas que incluem pelo menos a categoria gramatical, elas dizem respeito a construções gramaticais precisasou mesmo a empregos específicos das palavras. Por isso, o carácter exacto ou inexacto das restrições formalizadas permanece inalterado no caso da introdução de novas etiquetas lexicais que descrevam novas acepções das mesmas palavras no dicionário. Consideremos por exemplo a seguinte frase:

Ele está feliz

e a restrição gramatical que permite resolver a ambiguidade de género:

(31) Em <*ele*.PRO:3ms> <*estar*.V> <A:s>, o adjectivo deve estar no masculino.

Imaginemos agora o aparecimento de uma revista com o título *Ele*, e a introdução de uma etiqueta <*Ele*.Npr:fs> no dicionário de nomes próprios. A restrição gramatical (31) aplicada a:

(32) *Ele* (= *A revista Ele*) *está cheia de reportagens esta semana*

elimina correctamente a análise: <*ele*.PRO:3ms> <*estar*.V:P3s> <*cheio*.A:fs>. Não se aplica às análises <*Ele*.Npr:fs>, e portanto não as elimina. Se reformularmos a restrição (31), sem atender à convenção de formalização em discussão, obteremos:

(33) Em *ele* <*estar*.V> <A:s>, o adjectivo deve estar no masculino.

Com a introdução de <*Ele*.Npr:fs> no dicionário, esta versão revela-se errada: aplicada a (32), elimina a análise <*cheio*.A:fs>, conservando apenas, para esta forma, a descrição <*cheia*.N:fs>. A análise correcta é, portanto, eliminada.

Pode concluir-se que a convenção diminui o grau de interdependência dos conteúdos do dicionário e das gramáticas. Mais especificamente, o uso da versão (31) não evita a obrigatoriedade de introduzir a nova etiqueta no dicionário de nomes próprios, já que, de contrário, (31) eliminaria a análise correcta de *cheia* em (32); mas, uma vez que a nova etiqueta é suficientemente específica, não é necessário modificar (31) aquando da sua introdução.

Esta situação paradoxal, em que uma limitação do poder expressivo do formalismo facilita a utilização do próprio poder expressivo, é bem conhecida no domínio da construção de sistemas informáticos complexos. A programação orientada para objectos é um tipo de programação particularmente bem adaptada à criação e à evolução de sistemas complexos. Um dos meios utilizados eficazmente nesse quadro, para facilitar o trabalho de programação, é proibir o programador de usar determinadas variáveis em determinadas condições.

9.3. *Imbricação de zonas de aplicação*

Quando as restrições gramaticais são aplicadas a um texto, o reconhecimento das sequências descritas selecciona fragmentos do texto a que chamaremos zonas de aplicação. Por exemplo, a zona de aplicação de (33) abrange sempre três palavras. Os exemplos utilizados até agora neste capítulo, por serem exageradamente simples, reconhecem apenas sequências de duas ou três palavras, mas, na prática efectiva, a formalização de restrições gramaticais implica muitas vezes uma descrição de contextos mais amplos. Tomando em conta a densidade de ambiguidades lexicais, uma gramática de resolução de ambiguidades contém geralmente numerosas restrições, cujas zonas de aplicação nos textos apresentam imbricações, devido ao facto de os contextos relativos a elementos adjacentes se sobreporem parcialmente. A propósito das frases em (28), já evocámos a possibilidade de uma restrição eliminar a ambiguidade de *se*, e de outra utilizar o mesmo *se*, etiquetado como conjunção, como marca de limite de oração. Em tal situação, a forma *se* pertenceria às zonas de aplicação de ambas as restrições. Duas zonas de aplicação da mesma restrição podem mesmo apresentar tal imbricação, se o início e o fim do contexto descrito se sobrepuserem. Uma zona de aplicação pode mesmo estar completamente incluída noutra.

Em tais casos, o efeito da aplicação das restrições gramaticais depende da especificação associada ao formalismo. Duas situações são imagináveis e já foram testadas em sistemas reais: a aplicação de uma restrição a uma determinada zona pode depender, ou não, da presença eventual de uma imbricação com outra.

A primeira situação é a de sistemas que, em caso de imbricação entre duas zonas de aplicação, apenas verificam a restrição numa das duas zonas, a da esquerda, por exemplo. Com essa convenção, duas restrições gramaticais podem dar resultados correctos quando aplicadas separadamente, mas ser incompatíveis quando as zonas de aplicações acabam por se sobrepor. É até teoricamente possível que, quando se acrescenta uma restrição a uma gramática de filtragem existente, a taxa de redução de ambiguidades diminua, uma vez que impede a aplicação de restrições já presentes na gramática. Nesse tipo de situações, o resultado da aplicação de uma restrição gramatical depende não só do contexto nela descrito, mas também dos contextos de outras restrições que com ela possam estar em imbricação. A presença de imbricações alarga o contexto que determina o resultado da

aplicação, mudando, portanto, a interpretação da restrição gramatical em função das interacções entre zonas de aplicação adjacentes.

Na segunda situação, o próprio formalismo garante que o resultado da aplicação de cada restrição a cada zona de aplicação seja independente do resto do texto, tanto à esquerda como à direita. Isso não impossibilita a aparente "cooperação" entre regras, exemplificada a propósito das frases (28): a restrição gramatical que utiliza <*se*.CONJ> como marca de limite de oração não se aplica às análises de <*se*.PRO:3>, e, logo, não as elimina; elas são eliminadas de forma independente pela outra restrição, graças à presença do verbo à direita.

A segunda solução é mais confortável do ponto de vista da descrição das restrições gramaticais, uma vez que os efeitos de cada uma delas são independentes das outras: não são alteradas pela introdução de novas restrições, apenas se acrescentam aos efeitos das novas. Além disso, a extensão do contexto relevante a cada restrição permanece estritamente local, não se afastando da que foi decidida pelo autor da descrição.

9. Tipologia dos formalismos de resolução de ambiguidades

Já vimos que um formalismo de resolução de ambiguidades define simultaneamente as convenções de descrição das restrições gramaticais, e o resultado da aplicação das restrições às análises geradas pela consulta do dicionário. Além disso, estes dois aspectos estão intimamente relacionados.

A filtragem, por exemplo, pode ser concebida de forma positiva ou negativa, conforme se descrevem as análises que devem ser conservadas ou as que devem ser eliminadas. Esta opção diz respeito não só à descrição das restrições mas também à especificação do resultado da filtragem. As duas soluções são teoricamente possíveis, e foram já ambas realizadas em sistemas reais. Contudo, do ponto de vista da aplicação não apresentam as mesmas vantagens.

À luz das várias experiências realizadas nos últimos anos, apresentaremos uma tipologia dos formalismos de resolução de ambiguidades lexicais e uma avaliação do seu potencial respectivo. Classificá-los-emos em dois tipos a partir do seguinte critério de distinção. Consideraremos a operação automática elementar pela qual uma determinada análise de uma dada frase é eliminada ou conservada por aplicação de uma determinada regra descrita no formalismo. Tal decisão elementar depende das condições formalizadas na regra. Conceptualmente, toda a filtragem de análises pode ser vista como uma repetição deste mecanismo de decisão elementar. Dependendo do formalismo considerado, aquela operação pode ter em conta condições relacionadas com:

- o conjunto completo de análises concorrentes da frase, incluindo aquela a que a decisão diz respeito, ou
- apenas a análise à qual a decisão diz respeito, independentemente das outras.

Designaremos os formalismos do primeiro tipo como formalismos de filtragem dependente, e os do segundo como formalismos de filtragem independente. Existem na literatura exemplos destes dois tipos de formalismo. Pedimos desculpa ao leitor pela introdução dos termos algo estranhos «filtragem dependente e independente». Para designar o primeiro tipo, não existe ainda nenhum termo específico; quanto ao segundo formalismo, a designação proposta é *monótono* (Koskenniemi, 1990), que evoca o princípio geral da filtragem de análises por aplicação de restrições formais.

10.1. *Formalismos de filtragem dependente*

Conforme a definição que demos deste tipo de formalismo de resolução de ambiguidades lexicais, a decisão de eliminar ou conservar uma determinada análise de uma frase pode depender daquela análise, bem como de outras análises da mesma frase, ou pelo menos das que ainda não tenham sido eliminadas por aplicação de outras regras.

Vejamos um primeiro exemplo de restrição gramatical deste tipo:

(34) Em qualquer sequência com a forma <*ter*.V> <V.Kms>, todas as ambiguidades da primeira forma são resolvidas a favor do verbo *ter*, e todas as ambiguidades da segunda são resolvidas a favor do particípio.

Apliquemos esta restrição à seguinte frase:

(35) *Tendo consultado a evolução das cotações, decidiu comprar*

A etiqueta <*tender*,V.P1s> de *tendo* é correctamente eliminada. A filtragem é dependente já que a decisão acerca da análise <*tender*,V:P1s> é tomada devido à presença de outra etiqueta, <*ter*,V:G> para a mesma palavra, e, portanto, com base em propriedades de outras análises. Se uma outra regra tivesse previamente eliminado de forma incorrecta a etiqueta <*ter*,V:G>, a aplicação de (34) deixaria a frase inalterada, já que as condições de aplicação não seriam satisfeitas. Os formalismos de filtragem dependente permitem a expressão de restrições gramaticais com essa propriedade formal.

Noutros exemplos de restrições gramaticais do mesmo tipo, o modo de reconhecimento do contexto de aplicação tem em conta as ambiguidades:

(36) Quando uma forma ambígua entre <V:P3s> e <N:fs>, como *ajuda*, é seguida por um <A:fs>, a ambiguidade é resolvida a favor do <N:fs>.

Esta restrição aplicada, por exemplo, à frase:

Pediu uma ajuda financeira

elimina correctamente a etiqueta <*ajudar*,V:P3s>. A filtragem é dependente porque a condição para eliminar as análises em <*ajudar*,V:P3s> inclui a presença da etiqueta <N:fs>, que está associada à mesma forma, e pertence, portanto, a outras análises.

Enfim, um terceiro exemplo ilustra um modo de reconhecimento em que uma forma só é identificada se todas as suas ambiguidades já tiverem sido resolvidas. Repare-se que a restrição (34) dá resultados errados quando aplicada a:

(37) *Não acho o sentido desses termos apropriado à situação*

devido à etiqueta verbal de *termos*. Para melhorar (34), podemos imaginar outra versão:

(38) Em qualquer sequência com a forma <*ter*.V> <V.Kms>, se todas as ambiguidades da primeira forma já tiverem sido resolvidas a favor do verbo *ter*, então todas as ambiguidades da segunda são resolvidas a favor do particípio.

Esta versão conserva correctamente a análise correcta <*termo*.N:mp> em (37), excepto se outra restrição já tiver eliminado (incorrectamente) todas as análises em <*termo*.N:mp>. Todavia, não consegue resolver a ambiguidade de *tendo* em (35). Aqui, a filtragem é dependente porque a condição de aplicação de (38) inclui a ausência de análises das quais conste uma etiqueta lexical concorrente de <*ter*.V>.

Vários sistemas de resolução de ambiguidades lexicais são baseados em formalismos de filtragem dependente (Silberztein, 1994; Tzoukermann *et al.*, 1995); em Laval (1995: 103), um dos dois formalismos propostos é do mesmo tipo.

Antes de avaliar o potencial aplicativo deste tipo de formalismos, examinemos algumas propriedades formais dedutíveis da sua definição.

A primeira propriedade é a presença de interacções e dependências entre restrições incluídas na mesma gramática. Já que a decisão de eliminar ou conservar uma determinada análise de uma dada frase pode depender das outras análises da mesma frase que ainda não tenham sido eliminadas por aplicação de outras regras, o resultado da aplicação de uma restrição gramatical depende das outras restrições já aplicadas antes. Por outras palavras, a interpretação e os efeitos de uma restrição gramatical podem ser diferentes na ausência ou presença de outra restrição gramatical.

Essa primeira propriedade tem três consequências.

a) Quando se acrescentam novas restrições a uma gramática de filtragem existente, cujo desempenho é conhecido, a taxa de redução de ambiguidades lexicais pode diminuir em vez de aumentar, pois, se uma nova restrição se aplica ao texto antes de ter sido aplicada alguma da versão precedente, a gramática pode eliminar menos análises do que fazia anteriormente.

b) O objectivo de manter o silêncio em zero torna-se mais difícil de alcançar à medida que se acumulam restrições gramaticais. Quando se introduz uma nova restrição numa gramática existente, os efeitos das outras restrições podem mudar, a tal ponto que restrições que anteriormente não eliminavam análises correctas podem passar a eliminá-las.

c) É possível definir se a aplicação de uma restrição gramatical, ou de uma gramática completa, a uma dada frase, é correcta ou não: basta comparar o resultado com a análise linguística que se deseja conservar. Porém, no quadro de um formalismo de filtragem dependente, essa definição não pode ser alargada à própria restrição gramatical, porque o resultado da aplicação de uma dada restrição a várias frases depende das outras restrições aplicadas. Portanto, ainda que os efeitos da aplicação das restrições sejam explicitamente especificados no formalismo, não é possível definir uma noção de restrição gramatical correcta.

A segunda propriedade dos formalismos de filtragem dependente é a necessidade de definir um ou vários modos de combinação das restrições gramaticais. De facto, já verificámos que se deve definir o efeito da aplicação de uma restrição a uma frase; é igualmente necessário definir o efeito da aplicação de várias restrições formalizadas separadamente. São realizáveis dois tipos de modos de combinação:

a) Na combinação por composição, o resultado da aplicação de uma restrição é considerado como o material ao qual se aplica a restrição seguinte. Trata-se de uma composição de funções na acepção matemática do termo. O resultado final depende da ordem de composição, como é fácil verificar com os exemplos (34)-(35). Uma variante deste modo de combinação consiste em repetir a aplicação da série inteira de regras, na mesma ordem de cada vez, até que se chegue a uma série de aplicações sem nenhum efeito: nessa altura, o conjunto de análises obtido é designado por ponto fixo, uma vez que não é alterado por novas aplicações das mesmas regras. Matematicamente, há uma garantia de que um ponto fixo é alcançado num número finito de iterações, dado que todos os conjuntos de análises são finitos e só podem diminuir; porém, é fácil verificar que o ponto fixo depende da ordem de composição, isto é, o processo pode alcançar vários pontos fixos diferentes a partir do mesmo texto, dependendo da ordem de composição.

b) Também se podem definir modos de combinação simultânea. O mais simples é: o resultado da aplicação de um conjunto de regras a uma frase é definido como o conjunto de análises que seriam conservadas por todas as regras, se cada uma fosse aplicada ao mesmo conjunto de análises como se nenhuma das outras existisse. Quando todas as restrições são

aplicadas desta forma, todas as interacções entre regras desaparecem; torna-se então possível definir uma noção de restrição gramatical correcta, pois o formalismo garante que uma correcta acumulação de restrições é correcta, e a construção de uma gramática de resolução de ambiguidades lexicais por acumulação de restrições gramaticais torna-se mais simples e mais segura. De facto, as propriedades formais deste tipo de formalismo tornam-no parecido com os formalismos de filtragem independente.

Outros modos de combinação simultânea podem ser definidos a partir de convenções, determinando se uma análise eliminada por uma restrição e conservada por outra é eliminada ou conservada, em função de vários critérios, por exemplo, da disposição e da imbricação das zonas de aplicação, ou de uma rede de relações do tipo regra/excepção entre as restrições lexicais, como no caso das duas restrições seguintes:

(39) Em *de o*, *o* é pronome
(40) Excepção a (39): em *de o* <N:ms>... <V:inf>, *o* é determinante.

A restrição (39) aplica-se correctamente ao exemplo (41), e (40) ao exemplo (42):

(41) *Desistiu de o convencer*
(42) *A obra atrasou-se em virtude de o operário se ter acidentado*

Porém, (40) deve ser marcada como excepção a (39), porque (40) eliminaria incorrectamente a análise de (42), em que <*o*.DET:ms>. Estes modos de combinação simultânea introduzem novas interacções e relações de dependência entre restrições gramaticais, com as consequências antes mencionadas. No exemplo da rede de regras e excepções, a decisão sobre uma determinada análise tem em conta as condições relativas às outras regras aplicadas à mesma análise e marcadas como excepções. As relações de dependência entre restrições gramaticais tornam difícil a manutenção e a extensão da gramática, porque qualquer modificação ou novo elemento pode modificar o comportamento de todas as restrições ligadas por relações de dependência directa ou indirecta.

Qualquer que seja o modo de combinação das regras, é necessária uma especificação precisa e complexa. Quando tal especificação existe, é possível definir uma noção de sistema de restrições correcto, mas mesmo assim nem sempre é possível definir uma noção de restrição gramatical correcta. Além disso, os autores das gramáticas de resolução de ambiguidades devem assimilar essa especificação, porque a correcção das suas gramáticas depende dela. O risco de erros de descrição é importante, dado que a concepção das restrições gramaticais está baseada na intuição linguística, uma intuição que é coerente com o que os autores pensam que o sistema faz. Porém, se não tiverem assimilado correctamente o formalismo, as suas intuições de restrições gramaticais podem estar erradas. Aliás, esta é a principal razão pela qual a simplicidade de um formalismo é uma qualidade primordial.

10.2. *Formalismos de filtragem independente*

Segundo a nossa definição deste tipo de formalismo de resolução de ambiguidades lexicais, a decisão de eliminar ou conservar uma determinada análise de uma dada frase depende apenas daquela análise, e não de outras análises da mesma frase. Trata-se, portanto, de formalismos com menor poder expressivo, pois as condições de aplicação das restrições gramaticais só podem incluir as propriedades da própria análise a que a decisão diz respeito. O que designamos por análise, neste contexto, é uma sequência de etiquetas lexicais gerada pela consulta de um dicionário electrónico.

As seguintes restrições gramaticais (Roche, 1992) pertencem a esse tipo de filtragem:

(43) <*o*.DET> não é seguido por um verbo num tempo finito.
(44) <*o*.PRO> não é seguido por um substantivo.

Estas restrições eliminam correctamente <*o*.DET:fs> em (45) e <*o*.PRO:3fs> em (46):

(45) *Não a pronuncia correctamente*
(46) *A pronúncia não é a correcta*

Quando o verbo e o nome são homógrafos, (43) e (44) eliminam correctamente as análises em <*o*.DET:fs> <V:3s> e <*o*.PRO:3fs> <N:fs>:

Disse que a espera fora tranquila

As análises em <*o*.DET:fs> <N:fs> e <*o*.PRO:3fs> <V:3s> são conservadas. A filtragem é independente porque (43) só se aplica às análises com o verbo, e (44) só às análises com o substantivo, como se o processamento de cada análise fosse inteiramente independente.

Vários sistemas de resolução de ambiguidades lexicais são baseados em formalismos de filtragem independente (Roche, 1992; Voutilainen, 1994; Laporte, Monceaux, 1999); em Laval (1995: 101), um dos dois formalismos propostos é do mesmo tipo. Todos estes sistemas usam autómatos finitos e seguem as orientações gerais de M. Gross (1989) e Koskenniemi (1990).

A propriedade formal mais notável destes sistemas é que os efeitos da aplicação de uma restrição gramatical são independentes das outras restrições aplicadas ou aplicáveis antes ou depois dela, ou até ao mesmo tempo. Em consequência desta propriedade, pode-se aplicar um conjunto de restrições gramaticais, quer simultânea, quer sucessivamente, e em qualquer ordem, sem modificação do resultado, com o seguinte princípio: uma análise é conservada pela gramática de resolução de ambiguidades se e só se é conservada por cada uma das restrições gramaticais nela incluídas. Portanto, basta uma restrição para eliminar uma análise. De facto, se uma determinada análise for eliminada por um conjunto de restrições, isso implica que pelo menos uma delas eliminou a análise, qualquer que seja a ordem de aplicação das restrições. Na prática, se a aplicação de uma gramática de resolução de ambiguidades eliminar uma análise correcta, o que se deseja absolutamente evitar, basta aplicar separadamente cada uma das restrições da gramática para determinar as que estão erradas.

O efeito da aplicação de uma combinação de restrições gramaticais é, pois, simples de definir e não apresenta as variantes que observámos no caso dos formalismos de filtragem dependente. O ponto fixo da aplicação iterativa das regras, por exemplo, é alcançado logo na primeira aplicação, o que torna a noção de ponto fixo supérflua.

Além disso, as restrições gramaticais formalizadas separadamente são independentes, e não apresentam interacções. Isso facilita a construção de uma gramática de resolução de ambiguidades por acumulação de restrições. De facto, quando se acrescenta uma nova restrição a uma gramática existente, o próprio formalismo garante que a gramática continua a ser aplicada com os mesmos efeitos, e que os efeitos da aplicação da nova restrição se lhe adicionam. Especificamente, isso implica que, à medida que se introduzem novas restrições na gramática, a taxa de redução de ambiguidades lexicais pode aumentar mas nunca diminuir.

É possível definir uma noção de restrição gramatical correcta, verificando que todas as análises eliminadas pela restrição gramatical são incorrectas, por referência às análises linguísticas subjacentes em que se deseja basear o sistema. O modo de combinação das restrições gramaticais garante que um conjunto de restrições correctas é correcto.

A descrição de restrições gramaticais com um formalismo de filtragem independente torna-se ainda desejável por outra razão: os autores das descrições podem ignorar completamente as eventuais ambiguidades dos elementos descritos, pois os contextos

distribucionais e sintácticos dos elementos linguísticos são reconhecidos independentemente das outras construções que com eles podem ser confundidas, sendo as análises processadas como se fossem totalmente distintas.

Enfim, uma gramática de resolução de ambiguidades lexicais, expressa com um formalismo de filtragem independente, possui automaticamente duas outras aplicações importantes.

- Pode ser usada para detectar erros não lexicais: quando a gramática é correcta e elimina todas as análises de uma frase, isso significa que a frase não tem nenhuma análise correcta e que é, portanto, incorrecta. No caso de um formalismo de filtragem dependente, essa reutilização tem pouca probabilidade de funcionar, porque as análises incorrectas são frequentemente reconhecidas por intermédio de outras análises, correctas ou incorrectas, mas em que há ambiguidade entre elas. Ora, esta situação de ambiguidade não tem razões de existir no caso de erros não lexicais.

- Pode igualmente ser usada para resolver as homofonias do tipo *paço/passo*, que perturbam o reconhecimento da fala por ausência de razões fonéticas para escolher a ortografia certa. Por exemplo, se uma palavra for transcrita *paço/passo*, e se uma gramática de redução de ambiguidades lexicais estabelecer, por observação do contexto, que a palavra é um verbo, a hipótese *paço* pode ser eliminada. Mais uma vez, os formalismos de filtragem independente permitem essa reutilização porque são orientados para a discriminação entre análises correctas e incorrectas, quaisquer que sejam as ambiguidades entre as formas; a mesma reutilização é muito menos realista com formalismos de filtragem dependente, dado que orientam a descrição em função de conjuntos de formas homógrafas, que são diferentes dos conjuntos de formas homófonas.

Conclusão

A resolução de ambiguidades lexicais é um objectivo simples à primeira vista, mas, de acordo com os exemplos acima, parece evidente que podem esperar-se resultados mais coerentes, fiáveis e utilizáveis do que aqueles considerados hoje como normais. Tal progresso não poderá ser alcançado se não se tiverem em conta análises linguísticas muito mais elaboradas: o conteúdo informativo dos dados linguísticos é tão importante neste domínio quanto o é nas outras áreas da linguística informática. Para além disso, esperamos que tenha ficado claro que também é absolutamente necessária uma reflexão que tome em conta os aspectos formais e abstractos do problema, a fim de que se possam conceber ferramentas suficientemente simples, que tornem possível a elaboração de boas gramáticas de resolução de ambiguidades.

REFERÊNCIAS

Benello, J.; A.W. Mackie, J.A. Anderson (1989), «Syntactic category disambiguation with neural networks», *Computer Speech and Language* 3 (pp. 203-217).

Brill, Eric (1992), «A simple rule-based part-of-dpeech tagger», in *Proceedings of the 3th Conference on applied Natural Language Processing*, Trento, Itália (pp. 152-155).

Carvalho, Paula (em preparação), *Gramáticas de resolução de ambiguidades no interior de grupos nominais*, Tese de Mestrado, FLUL.

Garrigues, Mylène (1997), «Une méthode de désambiguïsation locale nom/adjectif pour l'analyse automatique de textes», *Langages* 126, *La description syntaxique des adjectifs pour les traitements informatiques*, org. Nam Jee-sun, Paris: Larousse (pp. 60-78).

Greene, Barbara B.; G.M. Rubin (1971), *Automated grammatical tagging of English*. Providence, Rhode Island.

Gross, Maurice (1989), «The use of finite automata in the lexical representation of natural language», in *Electronics Dictionaries and Automata in Computational Linguistics*, orgs. M. Gross e D. Perrin, *Lecture Notes in Computer Science* 377, Berlim: Springer (pp. 34 – 50).

Gross, Maurice (1994), «Constructing Lexicon-Grammars», in *Computational approaches to the Lexicon*, orgs. B. T. S. Atkins e A. Zampolli, Oxford University Press (pp. 213-163).

Harris, Zellig S. (1962), *String Analysis of Language Structure*. Mouton.

Klein, Sheldon; R.F. Simmons (1963), «A Computational approach to grammatical coding of English words», *JACM* 10 (p. 334-337).

Koskenniemi, Kimmo (1990), «Finite-state parsing and disambiguation», in *Proceedings of COLING-1990*, org. H. Karlgren, Helsinki (pp. 229-232).

Laporte, Éric; A. Monceaux (1999), «Elimination of lexical ambiguities by grammars: the ELAG system», *Lingvisticae Investigationes* XXII, Amsterdam/Philadelphia: Benjamins (pp. 341-367).

Laval, Philippe (1995), «Un système simple de levée des homographies», *Lingvisticae Investigationes* XIX(1), Amsterdam/Philadelphie: Benjamins (pp. 97-105).

Leech, Geoffrey; R. Garside; M. Bryant (1994), «CLAWS4: The Tagging of the British National Corpus», in *Proceedings of COLING-94*, Kyoto (pp. 622-628).

Marcus, Mitchell P.; B. Santorini; M. Marcinkiewicz (1993), «Building a large annotated corpus of English: the Penn Treebank», *Computational Linguistics* 19(2) (pp. 315-330).

Marshall, Ian (1983), «Choice of grammatical word-class without global syntactic analysis: Tagging words in the LOB Corpus», *Computers in the Humanities* 17 (pp. 139-150).

Merialdo, Bernard (1994), «Tagging English text with a probabilistic model», *Computational Linguistics* 20(2) (pp. 155-171).

Milne, Robert (1986), «Resolving lexical ambiguity in a deterministic parser», *Computational Linguistics* 12(1) (pp. 1-12).

Ranchhod, Elisabete (1989), «Predicative nouns and negation», *Lingvisticae Investigationes* XIII(2), Amsterdam/Philadelphie: Benjamins (pp. 387-397).

Roche, Emmanuel (1992), «Text disambiguation by finite state automata, an algorithm and experiments on corpora», in *Proceedings of COLING-92*, Nantes.

Senellart, Jean (1999), *Outils de reconnaissance d'expressions linguistiques complexes dans de grands corpus*. Tese de doutorado, LADL, Universidade Paris 7, 290 p.

Silberztein, Max (1994), «INTEX: a corpus processing system», in *Proceedings of COLING-94*, Kyoto (pp. 579-582).

Tzoukermann, Evelyne; D.R. Radev; W.A. Gale (1995), «Combining linguistic knowledge and statistical learning in French part-of-speech tagging», in *Proceedings of the SIGDAT Workshop "From Texts to Tags: Issues in Multilanguage Analysis"*, European Chapter of the ACL, Dublin.

Voutilainen, Atro; P. Tapanainen (1993), «Ambiguity resolution in a reductionistic parser», in *Proceedings of the 6th Conference of the European Chapter of the ACL*, Utrecht (pp. 394-403).

Voutilainen, Atro (1994), «Morphological disambiguation», in *Constraint Grammar: a Language-Independent System for Parsing Unrestricted Text*, org. F. Karlsson, A. Voutilainen, J. Heikkilä, A. Attila, Berlin/New York: Mouton-de Gruyter.